CellPress

Molecular Cell
Article

Coregulation of Transcription Factor Binding and Nucleosome Occupancy through DNA Features of Mammalian Enhancers

Iros Barozzi,[1,3] Marta Simonatto,[1,3] Silvia Bonifacio,[1] Lin Yang,[2] Remo Rohs,[2] Serena Ghisletti,[1] and Gioacchino Natoli[1,*]
[1]Department of Experimental Oncology, European Institute of Oncology (IEO), Via Adamello 16, 20139 Milan, Italy
[2]Molecular and Computational Biology Program, University of Southern California, Los Angeles, CA 90089, USA
[3]Co-first author
*Correspondence: gioacchino.natoli@ieo.eu
http://dx.doi.org/10.1016/j.molcel.2014.04.006

SUMMARY

Transcription factors (TFs) preferentially bind sites contained in regions of computationally predicted high nucleosomal occupancy, suggesting that nucleosomes are gatekeepers of TF binding sites. However, because of their complexity mammalian genomes contain millions of randomly occurring, unbound TF consensus binding sites. We hypothesized that the information controlling nucleosome assembly may coincide with the information that enables TFs to bind *cis*-regulatory elements while ignoring randomly occurring sites. Hence, nucleosomes would selectively mask genomic sites that can be contacted by TFs and thus be potentially functional. The hematopoietic pioneer TF Pu.1 maintained nucleosome depletion at macrophage-specific enhancers that displayed a broad range of nucleosome occupancy in other cell types and in reconstituted chromatin. We identified a minimal set of DNA sequence and shape features that accurately predicted both Pu.1 binding and nucleosome occupancy genome-wide. These data reveal a basic organizational principle of mammalian *cis*-regulatory elements whereby TF recruitment and nucleosome deposition are controlled by overlapping DNA sequence features.

INTRODUCTION

The identification of histone marks (notably H3K4me1) and coregulators (such as the histone acetyltransferase p300) associated with functionally validated enhancers and/or with evolutionarily conserved, potential *cis*-regulatory sequences recently enabled the characterization of the genomic repertoire of candidate enhancers characteristic of distinct cell types (Heintzman et al., 2007). Terminally differentiated cells have a unique repertoire of enhancers (Creyghton et al., 2010; Heintzman et al., 2009; Rada-Iglesias et al., 2011; Stergachis et al., 2013; Visel et al., 2009) that is generated by transcription factors (TFs) that control lineage specification (Calo and Wysocka, 2013; Natoli, 2010). Specialized nucleosome-binding TFs (pioneer factors) (Zaret and Carroll, 2011) increase the local accessibility of nucleosomal DNA as determined by DNase I hypersensitivity assays (Thurman et al., 2012) and promote the deposition of enhancer-associated histone marks (Heintzman et al., 2007). Since most other TFs are unable to bind nucleosomal DNA, transcriptional regulation in a given cell type occurs almost exclusively within the accessible fraction of a much broader genomic repertoire of *cis*-regulatory elements. Therefore, the mutual interplay between nucleosomes, pioneer TFs, and TFs opportunistically binding to accessible DNA controls cell type-specific transcriptional outputs.

Factors that determine nucleosome occupancy can be broadly classified into three groups (Struhl and Segal, 2013): DNA sequence, *trans*-acting factors (including TFs and the transcriptional machinery), and chromatin remodeling enzymes. The role of DNA sequence in nucleosome occupancy has been the object of a long controversy (Struhl and Segal, 2013) centered on the relative role of nucleotide composition (Segal et al., 2006) versus DNA-bound barriers (Mavrich et al., 2008) and remodeler-driven nucleosome packing against barriers (Zhang et al., 2011) in determining nucleosome positioning in vivo. It is now clear that each of these mechanisms has a specific role in controlling nucleosomal organization and that sequence-driven nucleosome assembly can be overcome by *trans*-acting factors in specific instances and at specific locations.

The affinity of DNA sequences for the histone octamer spans several orders of magnitude, and computational models that use sequence features to predict nucleosome occupancy have been described (Ioshikhes et al., 2006; Kaplan et al., 2009; Segal et al., 2006; Tillo and Hughes, 2009). In particular, poly(dA:dT) tracts are stiff regions unable to bend around the histone octamer (Nelson et al., 1987; Suter et al., 2000), which accounts for nucleosome depletion at poly(dA:dT) sequences of $\geq$ 5 bp in length in *S. cerevisiae* gene promoters. In human cells, container sites (sequences able to generate positioned nucleosomes in in vitro assembly experiments) (Valouev et al., 2011) are demarcated by nucleosome-repelling poly(dA:dT) tracts flanking moderately (dG:dC)-rich, high-affinity regions for nucleosomes.

Both in yeast (Charoensawan et al., 2012) and in mammals (Gaffney et al., 2012; Lidor Nili et al., 2010), some TFs have been shown to contact genomic sequences encoding high

nucleosome occupancy. This is consistent with the notion that *cis*-regulatory elements, albeit nucleosome depleted in those cells in which they are bound by TFs, have an intrinsic propensity to be incorporated into nucleosomes (Tillo et al., 2010). These observations suggest that nucleosomes actively mask TF consensus sites but can be displaced by cooperatively binding TFs. However, although a trend linking sequence-encoded nucleosome occupancy and TF recruitment is detectable when analyzing genomic data (Tillo et al., 2010), this simple conceptual scheme does not explain the much higher complexity of the relationship between TF binding and sequence determinants controlling nucleosome occupancy.

The notion that nucleosomes restrict the access of TFs to the underlying regulatory DNA leads to a simple yet critical inference: in complex mammalian genomes, high nucleosome occupancy must be selectively encoded by sequences containing engaged TF binding sites, but not by the millions of randomly occurring sequences that are apparently identical to TF consensus binding sites but are not engaged by TFs (Pan et al., 2010). Clearly, while TF binding to a genomic site will not always and necessarily cause functional effects, consensus sequences that are not engaged are not likely to contribute to transcriptional control.

In this study, we have addressed the hypothesis that in mammalian genomes the information controlling the ability of TFs to recognize cognate sites in *cis*-regulatory elements (while ignoring randomly occurring consensus binding sites, heretofore indicated as nonfunctional sites) may at least partially coincide with the information that controls incorporation of the same sequence into nucleosomes. An attractive and biologically meaningful implication of this hypothesis is that conservation of nucleosome occupancy and TF binding sites would be subjected to the same evolutionary forces.

We used primary mouse macrophages in which the ETS family TF Pu.1, the master regulator of the myeloid lineage (Nerlov and Graf, 1998; Rosenbauer and Tenen, 2007; Scott et al., 1994), binds virtually the entire repertoire of H3K4me1-positive regions and a large fraction of transcription start sites (TSSs) (Ghisletti et al., 2010; Heinz et al., 2010). Pu.1 expression in fibroblasts (Ghisletti et al., 2010) or in Pu.1-negative myeloid precursors (Heinz et al., 2010) is sufficient to drive the deposition of H3K4me1 and to locally increase DNA accessibility. This suggests that Pu.1, together with other TFs expressed at different phases of myeloid differentiation (Lichtinger et al., 2012), may act as a pioneer factor to create the macrophage-specific repertoire of accessible *cis*-regulatory elements.

We found that Pu.1-bound consensus sites, but not those sites that are not bound in any of the Pu.1-expressing cell types, were shielded by nucleosomes in cells that do not express Pu.1. However, nucleosomes covering Pu.1 sites displayed a broad spectrum of occupancy and positioning, thus unveiling a complexity in the interplay between TF binding and nucleosome occupancy that was overlooked by previous analyses. Both at distal and TSS-proximal Pu.1 sites, nucleosome occupancy and positioning were encoded in the DNA sequence and could be recapitulated in vitro. We identified a minimal set of DNA features, including three-dimensional DNA shape, which discriminated bound from unbound Pu.1 consensus sites at genome scale with unprecedented accuracy for a mammalian TF. Critically, the same set of features predicted nucleosome occupancy of the same DNA elements with improved or similar performances compared to computational models specifically designed for this aim, thus suggesting that overlapping DNA sequence features control both nucleosome deposition and binding competence of Pu.1 consensus sites.

RESULTS

DNA Sequence Features Correlate with Distinct Nucleosome Profiles In Vivo

Mononucleosome-sized DNA fragments from limited micrococcal nuclease (MNase) digestion of mouse macrophage nuclei were subjected to paired-end sequencing. Each of the four biological replicates used was sequenced to generate ~200 million uniquely aligned, filtered, and properly paired sequence reads (Table S1). By pooling these four replicates, we obtained 825 million sequencing reads that allowed us to obtain a high-resolution view of nucleosome arrays. Sequencing reads centered on annotated TSSs generated a canonical asymmetric pattern with a nucleosome-depleted region (NDR) bracketed by the −1 and a more prominent +1 nucleosome (Figure S1A). Conversely, when TSS-distal Pu.1 peaks were used as central anchors, we detected regular arrays of nucleosomes (Figure 1A). Nucleosome depletion surrounding Pu.1-bound sites was independently observed in ChIP-seq experiments based on sonicated chromatin (Ghisletti et al., 2010; Heinz et al., 2010), thus suggesting that it was not due to digestion of labile nucleosomes with high sensitivity to MNase.

Pu.1 summit-centered nucleosome maps were ordered based on the decreasing occupancy of the central NDR and divided in deciles (Figure 1B). De novo motif discovery on the sequences from each decile returned as first hit the known Pu.1 binding site, with very similar statistical significance (Figure 1B, right). The median distance of the best motif match to the Pu.1 peak center was very similar in all deciles and comprised between 7 and 10 nt. Pu.1 binding scores were significantly higher in the first decile but similar across all the others (Figure 1C). Taken together, these results indicate that different Pu.1 binding affinities do not contribute to a different NDR occupancy. Considering a ± 1.5 kbp region centered on the Pu.1 peaks, the first decile showed a lower overall nucleosome occupancy than the tenth one (Figure 1D), indicating that differences in nucleosome organization extend beyond the central regulatory region. The two NDR-flanking nucleosomes were prominent in the tenth decile and almost absent in the first, thus contributing to the lower occupancy and to the apparently broader width of the NDR in this group. Therefore, qualitatively different classes of NDRs surrounding Pu.1 peaks could be identified, and these classes did not correlate with differences in Pu.1 occupancy. A representative snapshot is shown in Figure 1E.

Since RNA Polymerase II (Pol II) is associated with a subset of enhancers (De Santa et al., 2010; Kim et al., 2010; Koch et al., 2011), we analyzed its density in the deciles. Pol II reads showed higher density in the NDRs of higher deciles (Figure 1B). While this result suggests that Pol II did not contribute to the maintenance of the low-occupancy and broad NDR characteristic of

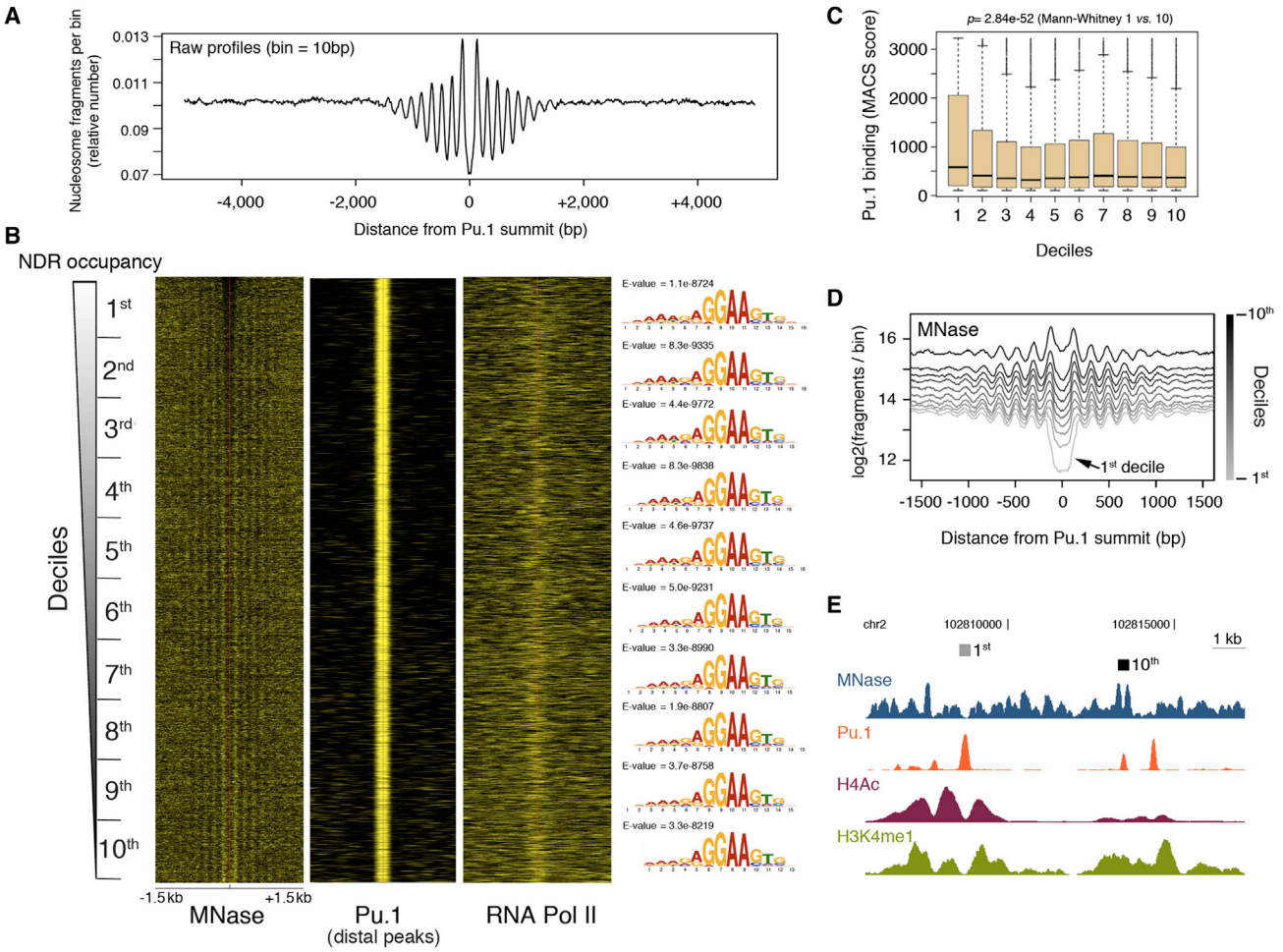

Figure 1. Regular Arrays of Nucleosomes Centered at Pu.1-Bound Enhancers in Macrophages

(A and B) Distribution of the midpoints of nucleosomal sequencing fragments centered on the summit of TSS-distal Pu.1 sites in macrophages (A). The number of fragments in each 10 bp bin was normalized by the total number of fragments in the area. The same information is shown in (B) as heatmap (first from the left), ordered from top to bottom based on decreasing occupancy of the NDR and divided in deciles. Heatmaps of Pu.1 and Pol II are also shown on the right of the MNase data. The counts exceeding the 95th percentile of the overall distribution were set to its value. Considering MNase data, these counts were then normalized in the range 0–1 separately for each region. The same procedure was applied to ChIP-seq data, except that the 0–1 normalization was applied to the entire data set. Sequence logos on the right show the Pu.1 binding motifs identified de novo in individual deciles and their *E*-values.

(C) ChIP-seq scores (MACS) of the Pu.1 peaks in in the different deciles.

(D) Distributions of the midpoints of the nucleosomal fragments at Pu.1-bound enhancers (divided in deciles according to B).

(E) A representative snapshot showing two NDRs of the first and the tenth decile. See also Figure S1.

the lower deciles, it may point to a role of Pol II in determining the occupancy and positioning properties of the higher deciles. However, depletion of the large Pol II subunit Rpb1 by a 4 hr alpha-amanitin treatment did not significantly alter nucleosome occupancy (M.S., I.B. and G.N., unpublished data). At TSS-proximal Pu.1 sites, the relationship between NDR occupancy and Pol II was opposite to that at enhancers, with higher Pol II loading in less occupied regions (Figures S1B and S1C).

We next analyzed the sequence features of the DNA associated with the distal Pu.1 binding sites. When considering the ensemble of all distal Pu.1-bound regions in macrophages, we detected features characteristic of nucleosome container sites (Valouev et al., 2011): an increase in the relative frequency of nucleosome-repelling AA dinucleotides and AAAA polynucleotides peaking at the −100 and +100 bp, with a central core of G+C-rich sequences that promote nucleosome occupancy (Tillo and Hughes, 2009) (Figure 2A). Next, we analyzed individual deciles separately. Consistent with the progressive increase in nucleosome occupancy, the G+C content increased from the first to the tenth decile ($p < 1 \times 10^{-15}$; Kruskal-Wallis test) (Figure 2B, left). Conversely, AA dinucleotides were more represented in the first decile and peaked at ± 100 nt (Figure 2C, left). In a reciprocal manner, the first decile showed a relative depletion of GC and CC dinucleotides in the flanks (Figure S2). Therefore, a signature of container sites was selectively found in the lower deciles.

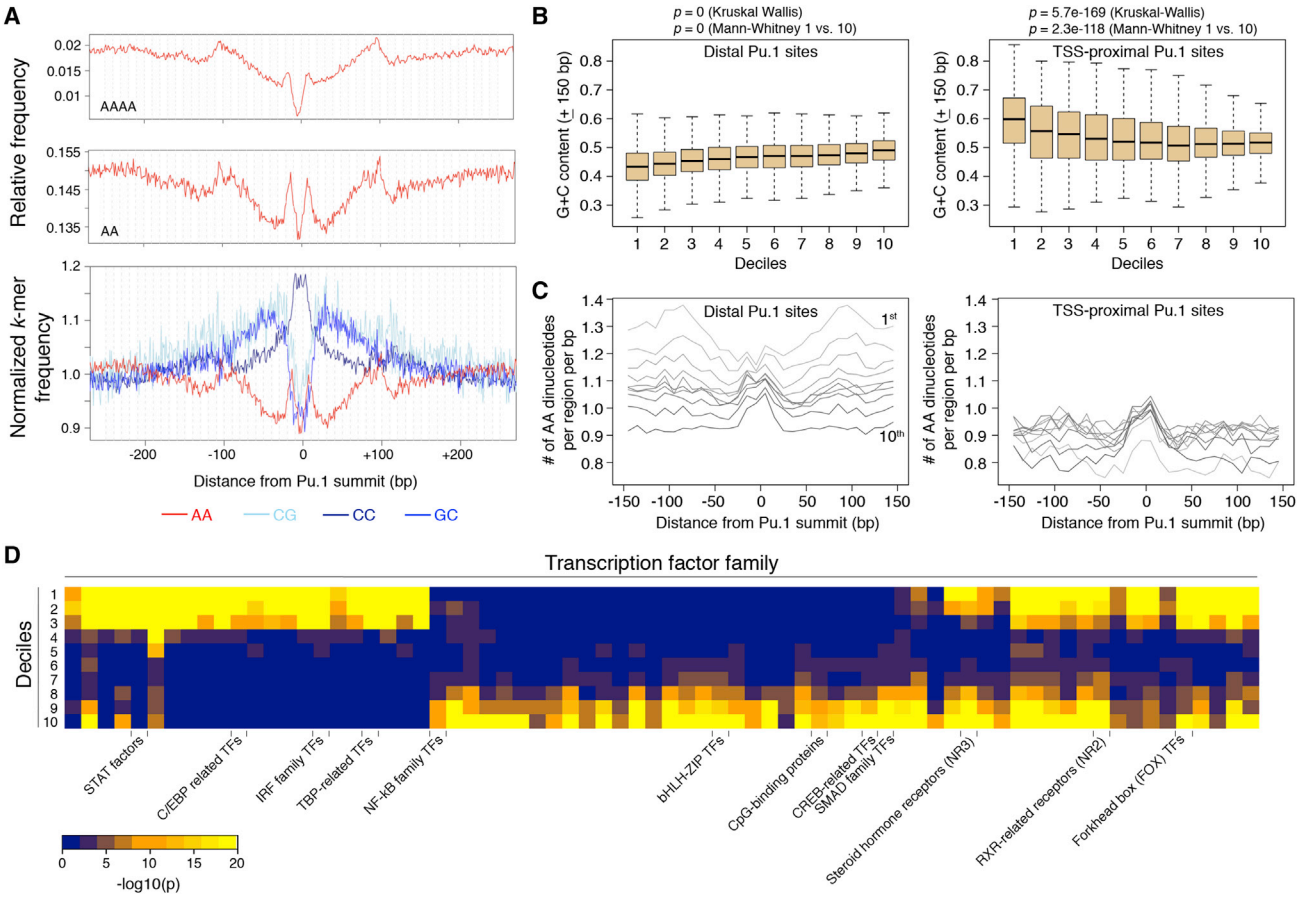

Figure 2. Sequence Features Discriminate among Enhancers with Different Nucleosome Occupancy and Positioning

(A) Distribution of AAAA tetranucleotides (top panel), AA dinucleotides (middle panel), and G/C-containing dinucleotides (bottom panel) are shown relative to the summit of TSS-distal Pu.1 peaks (the strong enrichment of CC/GG dinucleotides at the anchor point is enhanced by the central invariant nucleotides of the Pu.1 site, 5′-AGAGGAAGTG-3′).

(B and C) G+C content (B) and distribution of AA dinucleotides (C) in deciles at Pu.1-bound distal (left) and TSS-proximal (right) sites. See also Figure S2.

(D) Statistical overrepresentation of binding sites for TF families at Pu.1-bound distal sites divided in deciles (according to Figure 1B). For clarity, only selected TFs are indicated.

Compared to distal sites, sequence composition at TSS-proximal bound sites showed some fundamental differences. The G+C content at Pu.1-bound TSSs was much higher than the one at the distal sites (Figure 2B, right). At TSSs, the lower deciles (namely those with the lowest NDR occupancy; Figure S1C) showed the highest G+C content (Figure 2B, right). This is consistent with the notion that a very high G+C content (such as the one found at CpG islands) disfavors nucleosome assembly (Fenouil et al., 2012; Ramirez-Carrozzi et al., 2009). In fact, at TSSs the correlation between G+C content and deciles was inverted, with a progressive reduction from the first to the tenth decile. Therefore, the relationship between G+C content and nucleosome occupancy was bimodal, nucleosome occupancy being anticorrelated with both a very low G+C content (such as the one found at enhancers in the lower deciles) and a very high G+C content (such as the one found at promoters in the lower deciles). The AA frequency of the Pu.1-bound TSS-proximal regions was much lower than that observed at distal Pu.1 sites (Figure 2C, right). Moreover, TSS-proximal sites did not display the AA-rich flanks that delimit container sites and that could instead be detected in the lower deciles of the distal Pu.1-bound regions (Figure 2C, left).

Finally, we analyzed the co-occurrence of binding sites for other TFs at distal Pu.1 sites. Binding sites for some TF families (e.g., STAT and IRF) were overrepresented (relative to the distal Pu.1-bound sequences in their entirety) in the lower deciles, and some others (e.g., NF-κB and CREB) in the higher ones (Figure 2D).

Overall, these data indicate qualitative differences in sequence composition of the deciles that may directly impact both nucleosome assembly and cooperation with other TFs.

Selective Masking of Engaged Pu.1 Sites by Nucleosomes

To determine the impact of DNA sequence on nucleosomes at enhancers, we analyzed nucleosome occupancy in cell types that do not express Pu.1 (Figures 3A and 3B) and in in vitro reconstituted chromatin generated from recombinant histones

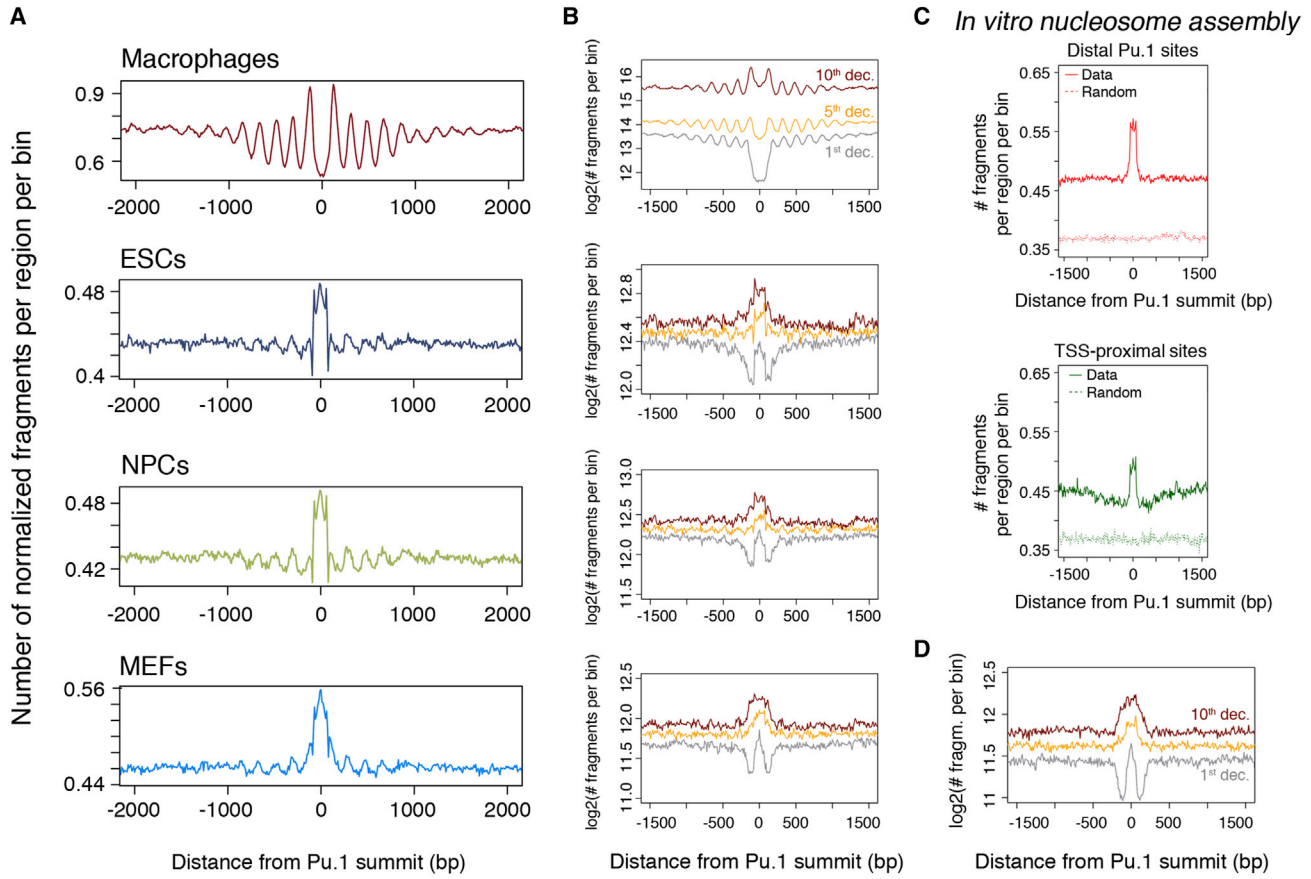

Figure 3. Pu.1-Bound, Nucleosome-Depleted Macrophage Enhancers Are Covered by Nucleosomes in Unrelated Cell Types and In Vitro

(A) Cumulative distributions of the midpoints of the nucleosomal fragments centered on distal Pu.1 sites in macrophages and in unrelated cells that do not express Pu.1 (ESCs, NPCs, and MEFs). The number of midpoints in each 10 bp bin was scaled according to the total number of regions and sequencing depth.

(B) Data were split in deciles (only the first, fifth, and tenth deciles are shown).

(C) Midpoint distributions from in vitro assembled nucleosomes. Data for distal and TSS-proximal sites are shown. See also Figure S3.

(D) MNase-seq data from in vitro nucleosomes divided in deciles.

(Figures 3C and 3D). Nucleosomal sequences from embryonic stem cells (ESCs), neural precursors (NPCs), and mouse embryonic fibroblasts (MEFs) (Teif et al., 2012) were aligned to the summit of Pu.1 peaks. In all three cell types, high nucleosome occupancy overlapping the macrophage Pu.1-bound nucleosome-depleted regions was detected (Figure 3A). Considering the deciles shown in Figure 1, the central NDR observed in the first decile in macrophages showed instead a focused nucleosomal signal bracketed by two narrow areas of nucleosome depletion in all other cells (Figure 3B). This result suggests a strongly positioned nucleosome controlled by container sites demarcated by AA-rich flanks (Figure 2C). The tenth decile was instead characterized by central nucleosomes with higher occupancy but weaker positioning.

Although nucleosome occupancy is affected by several factors, these data suggest that DNA sequence features contribute to promote nucleosome assembly at genomic regions that in macrophages are nucleosome-depleted due to Pu.1 binding. To directly define the role of DNA sequence in controlling the nucleosomal landscape at Pu.1 sites, we assembled nucleosomes in vitro. Assembly conditions in which DNA was not limiting were used to focus on the effects of the primary sequence on nucleosome assembly (Luger et al., 1999; Valouev et al., 2011). Sequencing data recapitulated previously reported features of in vitro assembled nucleosomes, such as the nucleosome depletion at CpG islands (Fenouil et al., 2012) that increased with CpG content but was less marked than that observed in vivo (Valouev et al., 2011) (Figure S3). The distribution of nucleosome midpoints (Figure 3C) indicates that genomic sequence features are sufficient to generate a focused increase in nucleosomal density at both distal and TSS-proximal sites bound by Pu.1 in macrophages. Consistent with the notion that formation of nucleosome arrays requires the activity of ATP-dependent remodelers (Zhang et al., 2011), we did not detect arrays in these conditions. When data were split according to the deciles shown in Figure 1, in vitro generated nucleosomes recapitulated the behaviors observed in cells that do not express Pu.1 (Figure 3D).

These data indicate that the DNA sequence encodes the deposition of nucleosomes at Pu.1-bound genomic sites. In cells

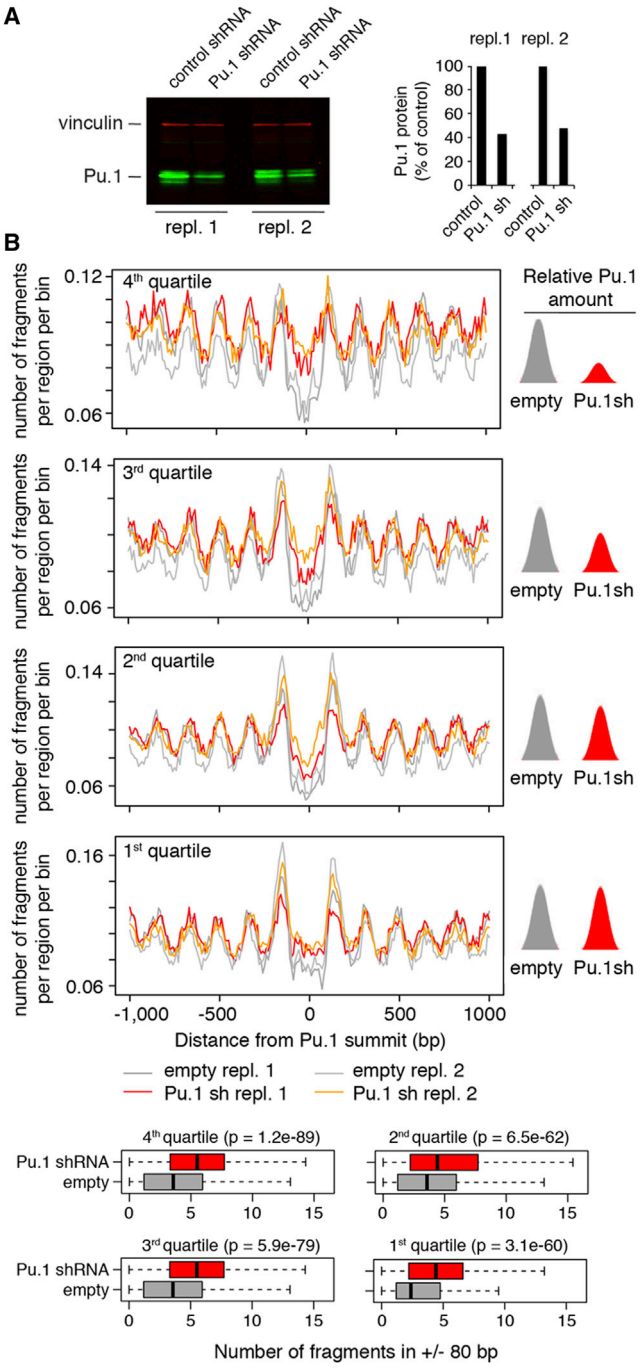

Figure 4. Effects of Pu.1 Depletion on Nucleosome Occupancy

(A) Acute depletion of Pu.1 in terminally differentiated macrophages using a retrovirus-encoded Tet-regulated shRNA. Data from two biological replicates are shown. Vinculin was used as loading control.

(B) Nucleosome occupancy in Pu.1-depleted macrophages. Pu.1 peaks were divided in quartiles based on the degree of signal reduction in Pu.1-depleted cells. The fourth quartile corresponds to Pu.1 peaks with the highest reduction in binding. Quartile-specific distributions of nucleosome fragment midpoints centered on the summit of Pu.1 peaks are shown. Midpoints found within ± 80 bp from the Pu.1 summit are summarized in the box plots at the bottom. For each quartile, the statistical significance of the difference is expressed by the p value of a Wilcoxon signed-rank test. See also Figure S4.

that do not express Pu.1, a range of behaviors was found. At one end (first decile), the regulatory region was covered by a strongly positioned nucleosome within an area of low nucleosome occupancy. In macrophages, the removal of this centrally positioned nucleosome resulted in the broad NDR characteristic of the first decile. At the opposite end (tenth decile), the sequence encoded higher occupancy but weak positioning. In macrophages, Pu.1 binding correlated with the partial displacement and repositioning of nucleosomes, thus resulting in a narrower NDR.

To address the role of Pu.1 in counteracting DNA sequence-driven nucleosome occupancy, we used an inducible retroviral vector to deplete Pu.1 from macrophages. Forty-eight hours after shRNA induction, a 60% depletion of Pu.1 was obtained in two independent experiments (repl. 1 and 2 in Figure 4A; notably, a complete depletion of Pu.1 would not be compatible with macrophage survival). We next carried out a Pu.1 ChIP-seq experiment to classify regulatory regions based on the level of reduction of Pu.1 binding, and we simultaneously analyzed nucleosome profiles. TSS-distal Pu.1 peaks identified by ChIP-seq were divided in quartiles based on the Pu.1 signal ratio in Pu.1-depleted versus control cells, the fourth quartile corresponding to the stronger reduction in Pu.1 binding. A strong and statistically significant ($p << 0.01$, Wilcoxon signed-rank test) increase in nucleosomal reads at Pu.1-bound enhancers was detected, particularly in the fourth quartile (Figure 4B, upper panel). Qualitatively similar results were obtained when considering the entire repertoire of Pu.1 peaks (Figure S4). Overall, these data indicate that genomic sites vacated by Pu.1 upon depletion tend to be reincorporated into nucleosomes.

To determine the ability of Pu.1 to bind different types of nucleosomal sites, we incubated in vitro assembled nucleosomes with macrophage nuclear extracts in order to allow the formation of protein-DNA complexes. Pu.1-bound nucleosomes were immunoprecipitated and sequenced. A Pu.1-immunodepleted nuclear extract was used as a reference (Figure 5A). Depending on the stringency, between 26% and 40% of the Pu.1 binding events observed in vivo were recapitulated in the in vitro assay (Figure 5B). When Pu.1 binding to in vitro assembled chromatin was analyzed considering the deciles shown in Figure 1, it became clear that while Pu.1 was able to strongly interact with sites in the first decile, it was less efficient at binding sites in the tenth decile ($p = 1.84 \times 10^{-278}$, Kruskal-Wallis test; Figure 5C), which is consistent with the in vivo binding data (Figure 1C). A representative in vitro ChIP-seq snapshot is shown in Figure 5D. Figure S5 shows a genomic snapshot of in vivo and in vitro assembled nucleosomes.

Since the Pu.1 consensus binding sites in different deciles are virtually identical (Figure 1B, right), these data and the in vivo data (Figures 1B and 1C) suggest that a high level of nucleosome occupancy has a detrimental impact on Pu.1 binding. To further address this issue, we analyzed the impact of transferring a 10 nt Pu.1 site (with a 15 nt extension on both sides) from intermediate deciles into the sequence context of lower or higher deciles. Nucleosomes were then assembled in vitro onto these chimeric sequences, and Pu.1 binding was measured by ChIP. As shown in Figure 5E, upon transferring

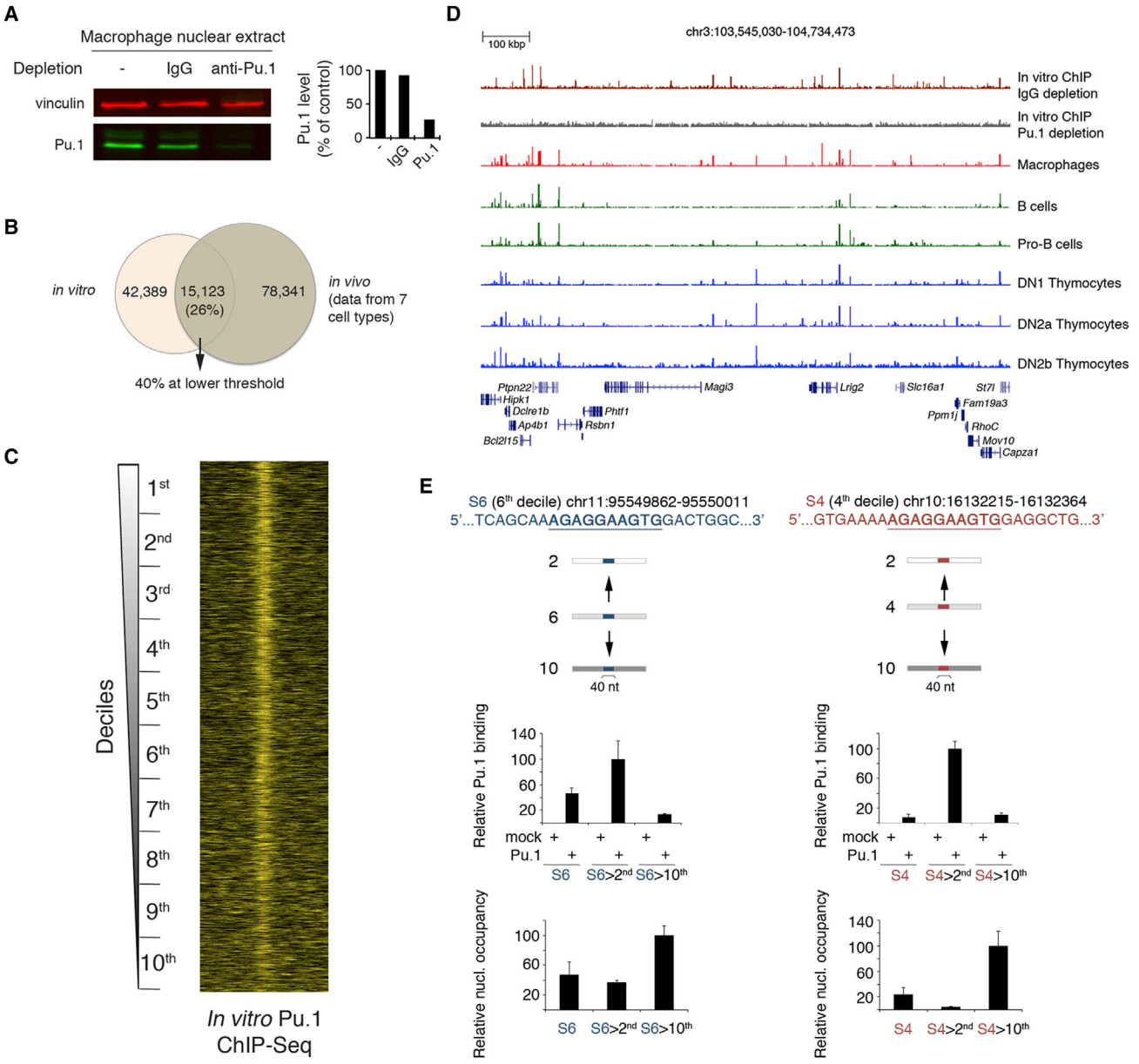

Figure 5. In Vitro Analysis of Pu.1 Binding to Nucleosomal DNA

(A) Pu.1 ChIP-seq on in vitro assembled nucleosomes. Macrophage nuclear lysates (used as the source of Pu.1) were incubated with in vitro assembled chromatin. Prior to incubation with nucleosomes, nuclear lysates were reacted twice either with control rabbit IgG or anti-Pu.1 antibody coupled to paramagnetic beads. Immunodepletion with anti-Pu.1 antibodies resulted in an almost complete loss of Pu.1 from lysates. Vinculin: loading control.

(B) Venn diagram showing the overlap between in vitro and in vivo Pu.1 binding.

(C) Heatmap of in vitro Pu.1 binding, showing the relative density of nucleosome midpoints. In vitro ChIP signals were sorted according to nucleosome occupancy in macrophages (Figure 1B). See also Figure S5.

(D) A representative ChIP-seq snapshot.

(E) Pu.1 sites and flanks (40 nt) from the sixth or fourth decile were transferred to sequences of higher (tenth) or lower (second) deciles and used for nucleosome assembly and ChIP-qPCR. Mock transfected extracts and extracts from HEK293 cells transfected with a Pu.1 expression vector were used as indicated. Data are expressed as mean ± SEM.

the binding site from its original context (fourth or sixth decile) to the one of a higher (tenth) decile, an increase in nucleosome occupancy was paralleled by a reduction in Pu.1 binding; the opposite was observed when the same sequences were moved into the context of a lower decile. Therefore, sequences that intrinsically favor nucleosome occupancy correlate with weaker Pu.1 binding in vivo and interfere with the association of Pu.1 with its binding site in vitro.

Usage of Pu.1 Binding Sites Correlates with Nucleosome Occupancy

Data shown above demonstrate that the DNA sequence promotes nucleosome assembly at regions containing Pu.1 consensus sites that are bound in vivo. However, randomly occurring nucleotide combinations in mammalian genomes lead to the casual generation of nonfunctional TF consensus sites. Randomly occurring sites outnumber TF consensus sites contained in functional *cis*-regulatory elements and bound by their cognate TF in vivo. We reasoned that the information that discriminates between these two groups of sites might be coupled with the information relevant for nucleosome assembly.

Pu.1 is expressed only in the hematopoietic system and specifically in myeloid cells and in B and early T lymphocytes. We collected seven high-quality ChIP-seq data sets from Pu.1-expressing cells (Heinz et al., 2010; Mullen et al., 2011; Ostuni et al., 2013; Zhang et al., 2012) (Table S2). Collectively, Pu.1 peaks from these data sets (n = 96,685) virtually represent the entire repertoire of genomic regulatory elements that can be bound by Pu.1 in mouse cells (Figure 6A). Of these peaks, 41,472 (42.9%) overlap a canonical high-affinity Pu.1 binding site, which differs from those bound by other ETS proteins (Wei et al., 2010). The reference mouse genome contains an additional 571,738 Pu.1 sites with a computationally predicted high affinity (for a total of 613,210 Pu.1 sites): even assuming that a fraction of them may be bound in conditions not recapitulated in the data sets we collected, it is clear that the vast majority of them are not bound in vivo.

We next reanalyzed MNase-seq data separately for different groups of Pu.1 binding sites. In macrophages, nucleosome arrays were very similar at Pu.1 peaks with or without the presence of a canonical binding site, while unbound sites were not associated with detectable nucleosome arrays (Figure 6B, upper panel). Pu.1-bound elements showed instead an increase in nucleosome occupancy over the binding site in ESCs and NPCs, irrespective of the presence or absence of a canonical binding site. Conversely, canonical sites that were not bound by Pu.1 in any of the cell types analyzed did not display any clear increase in occupancy over the flanking regions (Figure 6B). Therefore, the ability of a *cis*-regulatory region containing a Pu.1 site to bind Pu.1 in vivo correlated with its affinity for nucleosomes (Figure 6C).

Identification of DNA Sequence and Shape Determinants of Pu.1 Binding Site Occupancy

These data suggest that nucleosomes may selectively mask Pu.1 sites contained in *cis*-regulatory elements. Therefore, we set out to identify DNA features that discriminate bound from unbound Pu.1 sites and to test whether the same determinants predict nucleosome occupancy in cells that do not express Pu.1.

We considered the whole set of Pu.1 binding events at high-affinity sites (41,472) and randomly extracted the same number of regions from the unbound sites as negative set. We used Support Vector Machines (Cortes and Vapnik, 1995) to evaluate to what extent the local genomic sequence is informative for Pu.1 binding in vivo. In the case of bound regions, 995 DNA features were assessed in 300 bp windows aligned to the summit of the ChIP-seq peaks, and to the invariant GGAA core of the Pu.1 binding site in the case of the unbound ones. Features tested included: (1) position weight matrices (PWMs) describing known binding preferences for TFs, (2) nucleotide words of length 2 or 4 (*k*-mers), (3) G+C content, (4) the average theoretical nucleosome occupancy of the region calculated with a published algorithm (Kaplan et al., 2009), (5) the overlap with classes of repetitive elements, and (6) three-dimensional DNA shape features predicted for the 10 bp in the ETS core motif and for an additional 15 bp on each side of the core. DNA shape was shown to affect protein-DNA recognition (Rohs et al., 2009) and to improve the prediction of binding specificities for bHLH TFs in yeast (Gordân et al., 2013) and human (Yang et al., 2014) and for homeodomain TFs in mouse and *Drosophila* (Dror et al., 2014). Given this large amount of features, we devised a selection procedure (Guyon and Elisseeff, 2003) to identify the smallest set with the highest predictive power (Figure 6D).

We first assessed the prediction accuracy of the Pu.1 binding preferences as determined only by in vitro protein binding microarrays (Wei et al., 2010). Bound sites showed better FIMO p values (Figure 6E), indicating a higher median affinity of the target sites in these regions compared to the unbound sites ($p = 2.31 \times 10^{-294}$, Mann-Whitney test). Nevertheless, Pu.1 sequence preferences alone were poor predictors of binding, resulting in an average accuracy of 58.5% (Figure 6F). Instead, starting with the entire set of 995 features and through feature selection, we achieved an average accuracy of 78% (Figure 6F).

We then analyzed the contribution of individual groups of features to the prediction accuracy (Figure 6F). Theoretical nucleosome occupancy and G+C content had similar performances (accuracy of 59%–60%), consistent with the notion that G+C content is a proxy for nucleosome occupancy (Tillo and Hughes, 2009). The role of cooperativity in Pu.1 binding to genomic sites is demonstrated by the high prediction accuracy of PWMs for partner TFs (72.2%). Remarkably, a small number of DNA shape features (Zhou et al., 2013) alone achieved an average prediction accuracy of 71.9%. In the combined model, DNA shape features of the ETS core boundaries and the −2, −1, and +1 flanking nucleotides were systematically selected (Table S3). Of the four DNA shape features used in this study (minor groove width, roll, propeller twist, and helix twist), minor groove width and roll were the predominant structural determinants of Pu.1 binding (Table S3).

To directly assess the impact of DNA shape features on Pu.1 binding, we carried out competitive electrophoretic mobility shift assays (EMSAs). We used two labeled Pu.1 sites (10 nt flanked by 7 nt on both sides) and a panel of unlabeled competitors corresponding to high-affinity mouse genomic Pu.1 sites (Table S4). These sites were either unchanged or mutated in the two nucleotides upstream or/and downstream of the 10 nt core Pu.1 site. Mutations were designed to cause effects on DNA shape that would be detrimental for Pu.1 binding. While mutations either upstream or downstream of the Pu.1 site had a negative impact on the competition efficiency (downstream mutations being collectively more efficient than the upstream ones), their combination showed the most negative effect (Figures S6A and S6B).

Besides ETS family motifs, motifs for TFs that are known to cooperatively bind at Pu.1 sites (such as Fos/AP-1 and IRF family

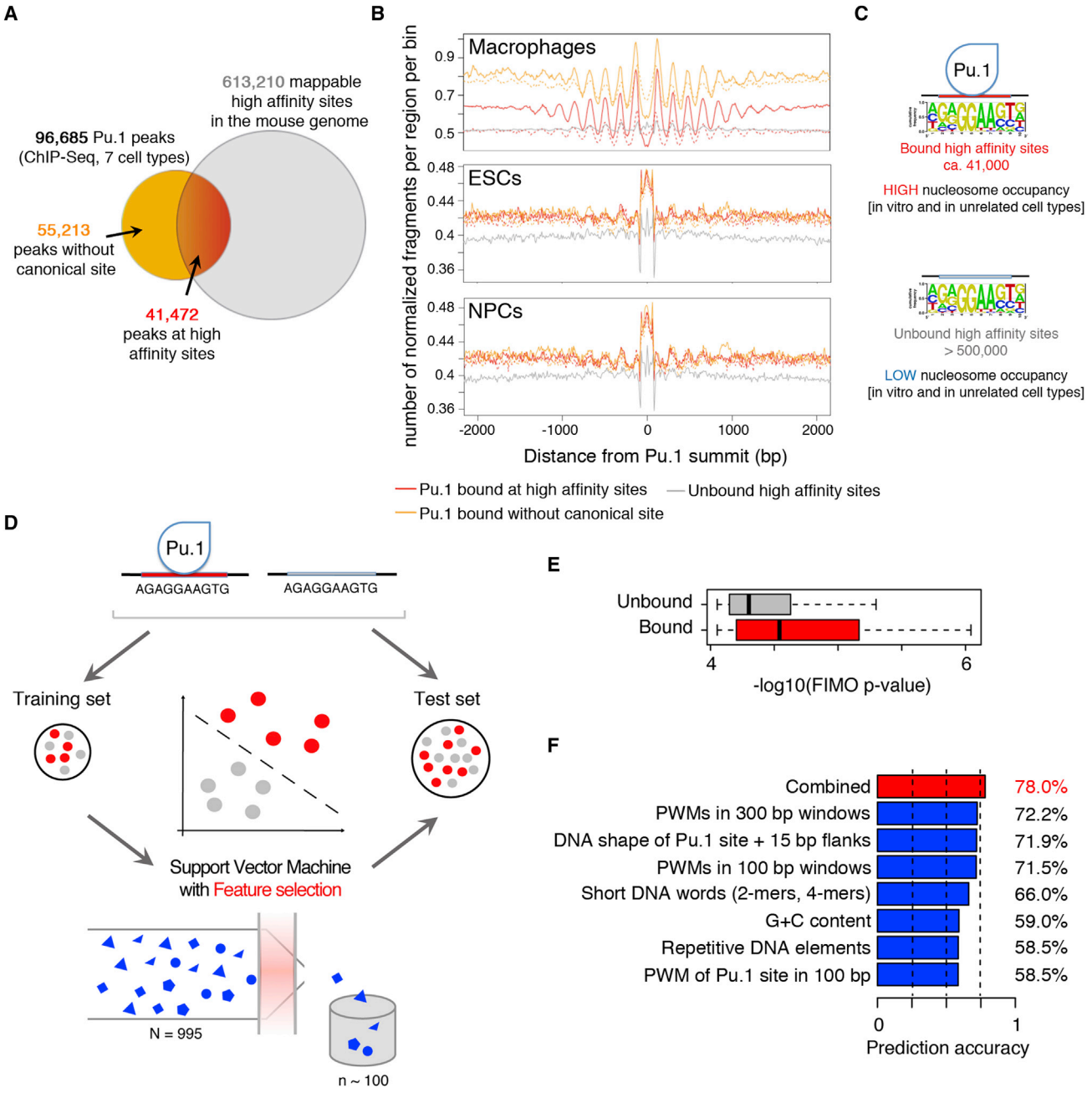

Figure 6. Pu.1 Binding Site Usage Correlates with Nucleosome Occupancy

(A) Venn diagram showing the overlap between Pu.1 peaks identified in ChIP-seq experiments from multiple cell types and computationally identified genomic Pu.1 sites.

(B) Cumulative distributions of nucleosome midpoints in macrophages, ESCs, and NPCs at Pu.1-bound high-affinity consensus sites (red), Pu.1-bound noncanonical sites (orange), and computationally identified consensus sites that are not bound in vivo (gray). Pu.1-engaged sites showing a low number of ChIP-seq tags in all cell types considered (see Supplemental Experimental Procedures) are indicated by dashed lines.

(C) Schematic representation of the relationship between Pu.1 binding and nucleosome occupancy.

(D) Schematic of the SVM approach used to predict in vivo engagement of Pu.1 sites and to identify their distinctive DNA sequence and shape features.

(E) Computationally predicted binding affinity of Pu.1 for bound and unbound genomic sites.

(F) Bar plots showing the prediction accuracies of the most predictive features selected by the SVM, divided in categories (blue) or all in combination (red). See also Figure S6.

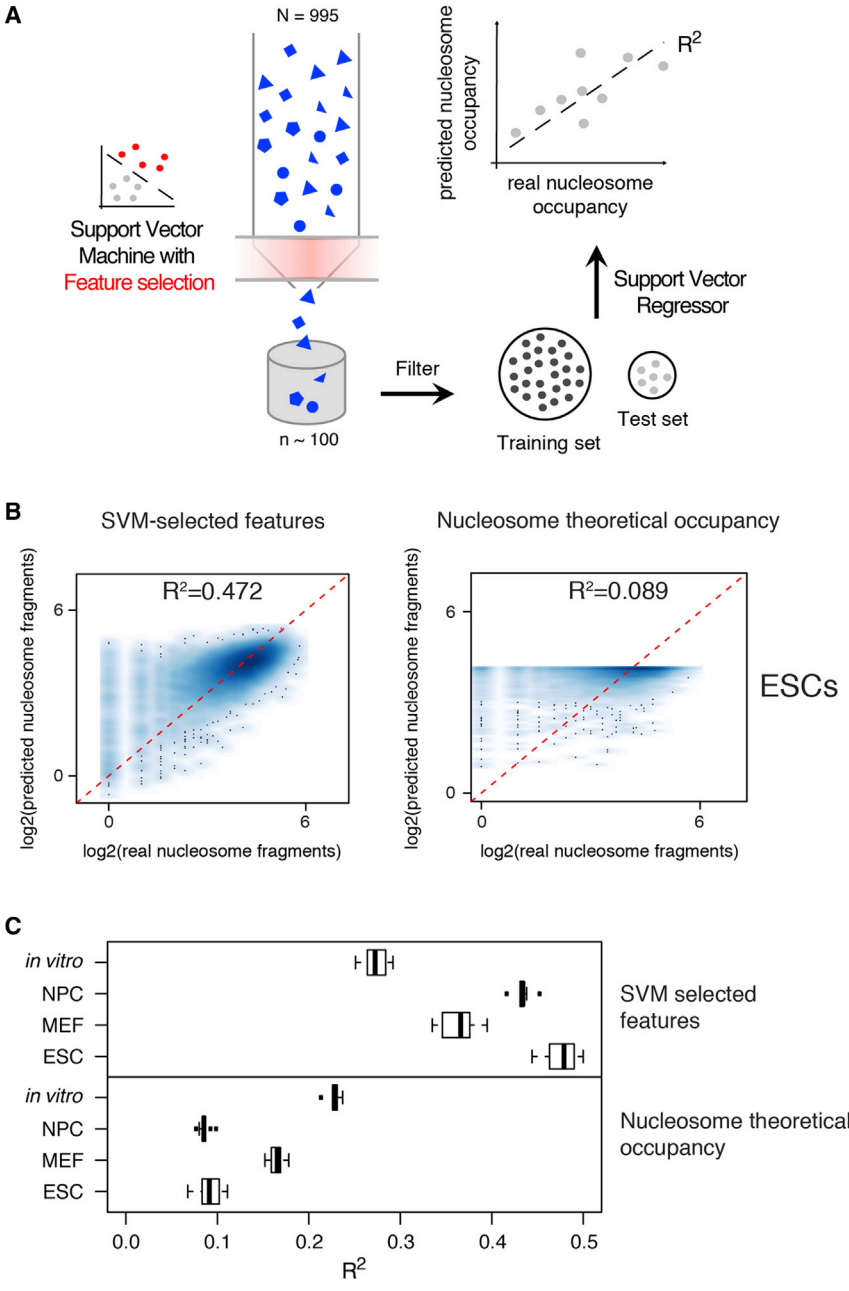

Figure 7. DNA Sequence and Shape Features Associated with Engaged TF Binding Sites Predict Nucleosome Occupancy

(A) Schematic representation of the SVR approach used to predict nucleosome occupancy using the DNA sequence and shape features that are predictive for Pu.1 binding.

(B) Smoothed scatterplots of the predicted vs. the observed log2-transformed values of nucleosome occupancy in ESCs at Pu.1 sites. The scatterplot on the right shows the results on the test data set using only theoretical nucleosome occupancy based on Kaplan et al. (2009). The one on the left shows the results using all the features selected by the SVM except nucleosome theoretical occupancy based on Kaplan et al. (2009).

(C) R2 values are robust to slight differences in the features (selected through multiple runs of the SVM) used as input for the SVR. Nucleosome theoretical occupancy is based on Kaplan et al. (2009).

TFs) (Ghisletti et al., 2010) were systematically selected. Moreover, we found the recurrent inclusion of G+C content/theoretical nucleosome occupancy as well as CG/GC/CC and AT/TA dinucleotides, which correlates with our previous observation that bound and unbound sites show a different potential to assemble nucleosomes when Pu.1 is not expressed.

Prediction of Nucleosome Occupancy Using Features that Predict Pu.1 Site Occupancy

We next asked whether the set of features that predict Pu.1 binding also predict nucleosomal patterns in cells that do not express Pu.1. The local nucleosome occupancy of regions containing Pu.1-bound sites was extracted from ESCs, NPCs, MEFs, and in vitro patterns. The number of nucleosome fragments spanning the center of each region was counted, and the log2-transformed value was used as a proxy for occupancy. The information for all the features except the theoretical nucleosome occupancy (Kaplan et al., 2009) was used to feed a support vector regressor (SVR) (Drucker et al., 1997), a variant of SVM for regression (Figure 7A). The set of bound and unbound sites was split into 90% training and 10% test data. The training data set was used to fit the experimentally determined nucleosome counts based on the sequence features. The model obtained was then used to predict the nucleosome counts over the test data set. Performance was evaluated through the coefficient of determination (R^2), calculated as the squared Pearson Correlation Coefficient among the predicted and the observed counts.

Results for a representative set of features are shown in Figure 7C as smoothed scatterplots of the predicted vs. the observed values. The features discriminating Pu.1-bound from unbound sites explained 47% of the variability in the nucleosome occupancy pattern at these sites in ESCs. Conversely, an SVR trained and tested using only the theoretical nucleosomes occupancy (Kaplan et al., 2009) explained less than 10% of the variability in the same data, which is in agreement with previous analyses (Tillo et al., 2010). These results were robust if slightly different sets of features (corresponding to multiple reinitializations of the SVM-based procedure used to predict Pu.1 binding) or different cell types were considered (Figure 7C). The results we obtained indicated improved or similar performance compared to previous

models specifically developed to predict nucleosome occupancy from the genomic sequence (Kaplan et al., 2009; Tillo and Hughes, 2009; van der Heijden et al., 2012).

Therefore, sequence determinants of Pu.1 binding could also encode part of the information for nucleosome affinity. Nevertheless, different DNA features are predicted to bring about quantitatively different effects on DNA binding and nucleosome occupancy. For instance, the DNA shape changes in the vicinity of the Pu.1 site should greatly impair Pu.1 binding without causing major effects on nucleosome assembly. Consistently, using the in vitro nucleosome assembly and Pu.1 ChIP assay described above, we found that mutations affecting DNA shape had a small but measurable detrimental effect on nucleosome assembly and a much higher impact on Pu.1 binding (Figure S6C).

DISCUSSION

The interplay between TF binding and nucleosome-mediated occlusion of the regulatory DNA sequences that TFs recognize is at the heart of regulated gene expression. However, understanding the determinants of this relationship has been hampered by some objective difficulties. First, engaged TF consensus sites are outnumbered by sites that, while characterized by an apparently identical affinity, represent nonfunctional, random nucleotide arrangements. Moreover, signals controlling nucleosome occupancy and positioning are degenerate and loose, and in vivo they can be overcome by DNA-bound barriers, thus complicating identification and dissection of *cis*- and *trans*-acting components in studies carried out in a single cell type. Finally, the huge number of nucleosomes associated with complex mammalian genomes imposes the requirement of a high sequencing depth (that until recently was unavailable or exceedingly expensive) to achieve the resolution required to faithfully measure their occupancy and positioning. The strategy used in this study allowed us to overcome some of these limitations and to reach a more advanced understanding of the basic properties of this essential regulatory relationship.

An important aspect of our strategy is that we anchored our analysis to a single TF, Pu.1, which pervasively marks enhancers in macrophages and imposes nucleosome depletion at these regulatory elements. Since Pu.1 is expressed exclusively in hematopoietic cells, and since its genomic distribution in Pu.1-expressing cell types has already been determined, it is possible to discriminate those randomly occurring Pu.1 sites that do not have binding competence from those that are contacted in vivo and are therefore potentially involved in transcriptional control. The classification of Pu.1 target sites based on their ability to bind Pu.1 in vivo allowed us to identify several molecular determinants of binding competence, to characterize different properties of binding sites in terms of their ability to drive nucleosome occupancy and positioning, and finally to determine whether and to what extent features controlling binding competence also affect nucleosome occupancy.

Collectively, Pu.1-bound *cis*-regulatory elements differed from unbound high-affinity Pu.1 sites in that, consistent with previous sequence-based predictions (Kaplan et al., 2009; Tillo and Hughes, 2009), they were preferentially associated with nucleosomes both in vitro and in vivo in unrelated cell types that do not express Pu.1. However, previous analyses missed the complexity of the relationship between nucleosomal occupancy and TF recruitment. In fact, when our MNase-seq data were deconvolved based on the occupancy of the central Pu.1-bound NDR, it became clear that the nucleosome-assembly ability of the genomic sequences bound by Pu.1 was not homogeneous. At one end of the spectrum, we found nucleosome container sequences (Valouev et al., 2011) able to drive the formation of a single, strongly positioned nucleosome within regions of overall lower nucleosome occupancy. At the other end we observed sequences determining a broad higher-occupancy context that extended on both sides of a centrally located, prominently but less strongly positioned nucleosome. Importantly, when analyzing Pu.1 recruitment to in vitro assembled chromatin, only the second configuration inhibited Pu.1 binding, thus suggesting that in spite of this broad nucleosome-mediated enforcement of Pu.1 sites, chromatin remodelers may be selectively required for full Pu.1 binding only to sequences characterized by an extended high nucleosomal occupancy.

An important issue relates to the functional impact of the different levels of intrinsic nucleosome occupancy and positioning in the deciles. In addition to the possibility that chromatin remodelers may be differentially required depending on the affinity of the underlying sequence for nucleosomes, lower and higher deciles were specifically enriched for binding sites of distinct TF families. For instance, binding sites for NF-κB, which controls the rapid induction of hundreds of inflammatory genes, showed their maximal relative enrichment in the higher deciles. Since NF-κB binding is greatly impaired by nucleosomes (Lone et al., 2013), the preferential inclusion of its binding sites within sequences that promote nucleosomal occupancy may provide the basis for a tight enforcement of its recruitment to regulatory sequences.

Another important aspect is that distal and TSS-proximal *cis*-regulatory elements bound by Pu.1 displayed fundamental differences in their sequence composition that resulted in distinct effects on nucleosome assembly. For instance, container sites were exclusively found at distal but not at TSS-proximal Pu.1 sites. Moreover, differently from distal sites, nucleosome depletion at TSS-proximal sequences with the lowest occupancy of the NDR was dependent on their high G+C content and not on Pu.1 occupancy.

Our data also demonstrate that the DNA sequence of *cis*-regulatory elements contains information that controls both binding competence of TF consensus sequences and nucleosome assembly. This co-occurrence explains the ability of the same DNA sequence features that discriminate bound from unbound Pu.1 sites to predict nucleosome occupancy in cells that do not express Pu.1. Whether such co-occurrence underlies direct causal relationships between DNA features that control TF recruitment (such as DNA shape characteristics) (Rohs et al., 2009) and nucleosome assembly remains to be determined. Moreover, the general relevance of this model outside of this specific set of regulatory sites will have to be assessed.

The notion that overlapping DNA sequence and shape features control both the ability of a genomic DNA sequence to recruit transcription factors and its propensity to be incorporated

into nucleosomes might have fundamental implications for both genomic biology and transcriptional control. First of all, it explains at the molecular level and at a genomic scale how regulatory elements can be selectively maintained under the gatekeeper activity of nucleosomes. Second, it implies that the same evolutionary forces that act to maintain the functionality of TF binding sites jointly control nucleosome deposition, thus preserving the gatekeeper function of nucleosomes during the evolution of regulatory DNA.

EXPERIMENTAL PROCEDURES

Nucleosome Mapping

A limited MNase digestion was carried out on intact macrophage nuclei to generate a mixture of mono- and polynucleosomes. Mononucleosomal DNA was isolated from agarose gels and used for library construction. A detailed description of the computational analyses is provided in the Supplemental Experimental Procedures.

ChIP Sequencing

ChIP was carried out starting from 5–8 × 10^6 cells, using a previously described protocol (Ghisletti et al., 2010).

In Vitro Nucleosome Assembly and In Vitro ChIP

Naked genomic DNA purified from mouse macrophages was sonicated to obtain fragments ranging from 600 to 2,000 bp. DNA was combined with recombinant histones (EpiMark Nucleosome Assembly Kit, NEB E5350) to generate nucleosomes by salt dialysis (Luger et al., 1999). In vitro assembled nucleosomes were digested with MNase and then incubated with macrophage-derived nuclear extracts to generate TF-nucleosome complexes. The in vitro ChIP-seq was carried out as described in the Supplemental Experimental Procedures.

Computational Methods

MNase-seq paired-end reads were mapped to the mouse genome using Bowtie (Langmead et al., 2009). Wiggle tracks at single base pair resolution were generated with BedTools (Quinlan and Hall, 2010). PeakSplitter (Salmon-Divon et al., 2010) was used to extract nucleosomal positions from this population-averaged profile. Paired-end fragments for ESCs, NPCs, and MEFs were retrieved from the literature (Teif et al., 2012).

ChIP-seq reads were aligned to the mouse genome using Bowtie, and peak calling was performed using MACS (Zhang et al., 2008). Peaks were annotated over Ensembl genes (Flicek et al., 2012). Pu.1-bound regions were sorted according to the NDR occupancy level. The number of midpoints of the nucleosomal fragments falling into the central 300 bp of each region was calculated and used as a proxy for the overall occupancy of the area. The genome-wide map of canonical Pu.1-binding sites (De Santa et al., 2010) was generated using FIMO (Creyghton et al., 2010).

Support vector machines (SVMs) (Cortes and Vapnik, 1995) were used to classify Pu.1-bound and unbound sites. Given a set of examples, an SVM training algorithm builds a model that can be used to categorize new examples. The LibSVM implementation (Chang and Lin, 2011) was used to train and test two-class SVMs. Given the large amount of features used, a selection procedure (Guyon and Elisseeff, 2003) to identify the smallest set with the highest predictive power was devised. Support vector regressors (SVRs) (Drucker et al., 1997) were applied to assess the fraction of variability in the nucleosomal occupancy patterns at Pu.1-bound and unbound sites that can be explained by the features selected by the SVM. DNA shape features were derived from a high-throughput data mining approach of all-atom Monte-Carlo predictions (Zhou et al., 2013). A detailed description of the computational methods and feature groups is provided in the Supplemental Experimental Procedures.

ACCESSION NUMBERS

Raw data sets are available for download at the Gene Expression Omnibus (GEO) database under the accession number GSE50762.

SUPPLEMENTAL INFORMATION

Supplemental Information includes six figures, Supplemental Experimental Procedures, and four tables and can be found with this article online at http://dx.doi.org/10.1016/j.molcel.2014.04.006.

AUTHOR CONTRIBUTIONS

I.B. and M.S. are equal contributors and are listed alphabetically. I.B. carried out most of the computational analyses in the paper; M.S. carried out most of the experimental work with the help of S.B. and S.G.; L.Y. and R.R. analyzed DNA shape features in the data sets; G.N. designed the study and wrote the paper with contributions from all authors.

ACKNOWLEDGMENTS

This work was supported by the European Research Council (ERC grant NORM to G.N.) and in part the National Institutes of Health (grants R01GM106056 and U01GM103804 to R.R.). R.R. is an Alfred P. Sloan Research Fellow. We thank B. Amati (IEO/IIT, Milan), J.C. Andrau (CIML, Marseille), and A. Agresti (HSR, Milan) for comments on the manuscript; M. Pelizzola, L. Riva, M. Morelli (IIT, Milan), and L. Fornasari (IFOM, Milan) for suggestions; L. Ferrarini (IFOM) for help with machine learning; and N. Habib, A. Weiner, and N. Friedman (HUJI, Jerusalem) for initial help with the SVM and the analysis.

Received: November 6, 2013
Revised: January 31, 2014
Accepted: March 19, 2014
Published: May 8, 2014

REFERENCES

Calo, E., and Wysocka, J. (2013). Modification of enhancer chromatin: what, how, and why? Mol. Cell *49*, 825–837.

Chang, C.-C., and Lin, C.-J. (2011). LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology *2*, 1–27.

Charoensawan, V., Janga, S.C., Bulyk, M.L., Babu, M.M., and Teichmann, S.A. (2012). DNA sequence preferences of transcriptional activators correlate more strongly than repressors with nucleosomes. Mol. Cell *47*, 183–192.

Cortes, C., and Vapnik, V. (1995). Support-vector networks. Machine Learning *20*, 273–297.

Creyghton, M.P., Cheng, A.W., Welstead, G.G., Kooistra, T., Carey, B.W., Steine, E.J., Hanna, J., Lodato, M.A., Frampton, G.M., Sharp, P.A., et al. (2010). Histone H3K27ac separates active from poised enhancers and predicts developmental state. Proc. Natl. Acad. Sci. USA *107*, 21931–21936.

De Santa, F., Barozzi, I., Mietton, F., Ghisletti, S., Polletti, S., Tusi, B.K., Muller, H., Ragoussis, J., Wei, C.L., and Natoli, G. (2010). A large fraction of extragenic RNA pol II transcription sites overlap enhancers. PLoS Biol. *8*, e1000384.

Dror, I., Zhou, T., Mandel-Gutfreund, Y., and Rohs, R. (2014). Covariation between homeodomain transcription factors and the shape of their DNA binding sites. Nucleic Acids Res. *42*, 430–441.

Drucker, H., Burges, C.J.C., Kaufman, L., Smola, A., and Vapnik, V. (1997). Support vector regression machines. In Advances in Neural Information Processing Systems, M. Mozer, M. Jordan, and T. Petsche, eds. (Cambridge, MA: MIT Press), pp. 155–161.

Fenouil, R., Cauchy, P., Koch, F., Descostes, N., Cabeza, J.Z., Innocenti, C., Ferrier, P., Spicuglia, S., Gut, M., Gut, I., and Andrau, J.C. (2012). CpG islands and GC content dictate nucleosome depletion in a transcription-independent manner at mammalian promoters. Genome Res. *22*, 2399–2408.

Flicek, P., Amode, M.R., Barrell, D., Beal, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fairley, S., Fitzgerald, S., et al. (2012). Ensembl 2012. Nucleic Acids Res. *40* (Database issue), D84–D90.

Gaffney, D.J., McVicker, G., Pai, A.A., Fondufe-Mittendorf, Y.N., Lewellen, N., Michelini, K., Widom, J., Gilad, Y., and Pritchard, J.K. (2012). Controls of nucleosome positioning in the human genome. PLoS Genet. *8*, e1003036.

Ghisletti, S., Barozzi, I., Mietton, F., Polletti, S., De Santa, F., Venturini, E., Gregory, L., Lonie, L., Chew, A., Wei, C.L., et al. (2010). Identification and characterization of enhancers controlling the inflammatory gene expression program in macrophages. Immunity *32*, 317–328.

Gordân, R., Shen, N., Dror, I., Zhou, T., Horton, J., Rohs, R., and Bulyk, M.L. (2013). Genomic regions flanking E-box binding sites influence DNA binding specificity of bHLH transcription factors through DNA shape. Cell Rep *3*, 1093–1104.

Guyon, I., and Elisseeff, A. (2003). An introduction to variable and feature selection. The Journal of Machine Learning Research *3*, 1157–1182.

Heintzman, N.D., Stuart, R.K., Hon, G., Fu, Y., Ching, C.W., Hawkins, R.D., Barrera, L.O., Van Calcar, S., Qu, C., Ching, K.A., et al. (2007). Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. Nat. Genet. *39*, 311–318.

Heintzman, N.D., Hon, G.C., Hawkins, R.D., Kheradpour, P., Stark, A., Harp, L.F., Ye, Z., Lee, L.K., Stuart, R.K., Ching, C.W., et al. (2009). Histone modifications at human enhancers reflect global cell-type-specific gene expression. Nature *459*, 108–112.

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., Cheng, J.X., Murre, C., Singh, H., and Glass, C.K. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol. Cell *38*, 576–589.

Ioshikhes, I.P., Albert, I., Zanton, S.J., and Pugh, B.F. (2006). Nucleosome positions predicted through comparative genomics. Nat. Genet. *38*, 1210–1215.

Kaplan, N., Moore, I.K., Fondufe-Mittendorf, Y., Gossett, A.J., Tillo, D., Field, Y., LeProust, E.M., Hughes, T.R., Lieb, J.D., Widom, J., and Segal, E. (2009). The DNA-encoded nucleosome organization of a eukaryotic genome. Nature *458*, 362–366.

Kim, T.K., Hemberg, M., Gray, J.M., Costa, A.M., Bear, D.M., Wu, J., Harmin, D.A., Laptewicz, M., Barbara-Haley, K., Kuersten, S., et al. (2010). Widespread transcription at neuronal activity-regulated enhancers. Nature *465*, 182–187.

Koch, F., Fenouil, R., Gut, M., Cauchy, P., Albert, T.K., Zacarias-Cabeza, J., Spicuglia, S., de la Chapelle, A.L., Heidemann, M., Hintermair, C., et al. (2011). Transcription initiation platforms and GTF recruitment at tissue-specific enhancers and promoters. Nat. Struct. Mol. Biol. *18*, 956–963.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. *10*, R25.

Lichtinger, M., Ingram, R., Hannah, R., Müller, D., Clarke, D., Assi, S.A., Lie-A-Ling, M., Noailles, L., Vijayabaskar, M.S., Wu, M., et al. (2012). RUNX1 reshapes the epigenetic landscape at the onset of haematopoiesis. EMBO J. *31*, 4318–4333.

Lidor Nili, E., Field, Y., Lubling, Y., Widom, J., Oren, M., and Segal, E. (2010). p53 binds preferentially to genomic regions with high DNA-encoded nucleosome occupancy. Genome Res. *20*, 1361–1368.

Lone, I.N., Shukla, M.S., Charles Richard, J.L., Peshev, Z.Y., Dimitrov, S., and Angelov, D. (2013). Binding of NF-κB to nucleosomes: effect of translational positioning, nucleosome remodeling and linker histone H1. PLoS Genet. *9*, e1003830.

Luger, K., Rechsteiner, T.J., and Richmond, T.J. (1999). Preparation of nucleosome core particle from recombinant histones. Methods Enzymol. *304*, 3–19.

Mavrich, T.N., Ioshikhes, I.P., Venters, B.J., Jiang, C., Tomsho, L.P., Qi, J., Schuster, S.C., Albert, I., and Pugh, B.F. (2008). A barrier nucleosome model for statistical positioning of nucleosomes throughout the yeast genome. Genome Res. *18*, 1073–1083.

Mullen, A.C., Orlando, D.A., Newman, J.J., Lovén, J., Kumar, R.M., Bilodeau, S., Reddy, J., Guenther, M.G., DeKoter, R.P., and Young, R.A. (2011). Master transcription factors determine cell-type-specific responses to TGF-β signaling. Cell *147*, 565–576.

Natoli, G. (2010). Maintaining cell identity through global control of genomic organization. Immunity *33*, 12–24.

Nelson, H.C., Finch, J.T., Luisi, B.F., and Klug, A. (1987). The structure of an oligo(dA).oligo(dT) tract and its biological implications. Nature *330*, 221–226.

Nerlov, C., and Graf, T. (1998). PU.1 induces myeloid lineage commitment in multipotent hematopoietic progenitors. Genes Dev. *12*, 2403–2412.

Ostuni, R., Piccolo, V., Barozzi, I., Polletti, S., Termanini, A., Bonifacio, S., Curina, A., Prosperini, E., Ghisletti, S., and Natoli, G. (2013). Latent enhancers activated by stimulation in differentiated cells. Cell *152*, 157–171.

Pan, Y., Tsai, C.-J., Ma, B., and Nussinov, R. (2010). Mechanisms of transcription factor selectivity. Trends Genet. *26*, 75–83.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842.

Rada-Iglesias, A., Bajpai, R., Swigut, T., Brugmann, S.A., Flynn, R.A., and Wysocka, J. (2011). A unique chromatin signature uncovers early developmental enhancers in humans. Nature *470*, 279–283.

Ramirez-Carrozzi, V.R., Braas, D., Bhatt, D.M., Cheng, C.S., Hong, C., Doty, K.R., Black, J.C., Hoffmann, A., Carey, M., and Smale, S.T. (2009). A unifying model for the selective regulation of inducible transcription by CpG islands and nucleosome remodeling. Cell *138*, 114–128.

Rohs, R., West, S.M., Sosinsky, A., Liu, P., Mann, R.S., and Honig, B. (2009). The role of DNA shape in protein-DNA recognition. Nature *461*, 1248–1253.

Rosenbauer, F., and Tenen, D.G. (2007). Transcription factors in myeloid development: balancing differentiation with transformation. Nat. Rev. Immunol. *7*, 105–117.

Salmon-Divon, M., Dvinge, H., Tammoja, K., and Bertone, P. (2010). PeakAnalyzer: genome-wide annotation of chromatin binding and modification loci. BMC Bioinformatics *11*, 415.

Scott, E.W., Simon, M.C., Anastasi, J., and Singh, H. (1994). Requirement of transcription factor PU.1 in the development of multiple hematopoietic lineages. Science *265*, 1573–1577.

Segal, E., Fondufe-Mittendorf, Y., Chen, L., Thåström, A., Field, Y., Moore, I.K., Wang, J.-P.Z., and Widom, J. (2006). A genomic code for nucleosome positioning. Nature *442*, 772–778.

Stergachis, A.B., Neph, S., Reynolds, A., Humbert, R., Miller, B., Paige, S.L., Vernot, B., Cheng, J.B., Thurman, R.E., Sandstrom, R., et al. (2013). Developmental fate and cellular maturity encoded in human regulatory DNA landscapes. Cell *154*, 888–903.

Struhl, K., and Segal, E. (2013). Determinants of nucleosome positioning. Nat. Struct. Mol. Biol. *20*, 267–273.

Suter, B., Schnappauf, G., and Thoma, F. (2000). Poly(dA.dT) sequences exist as rigid DNA structures in nucleosome-free yeast promoters in vivo. Nucleic Acids Res. *28*, 4083–4089.

Teif, V.B., Vainshtein, Y., Caudron-Herger, M.Ø., Mallm, J.-P., Marth, C., Höfer, T., and Rippe, K. (2012). Genome-wide nucleosome positioning during embryonic stem cell development. Nat. Struct. Mol. Biol. *19*, 1185–1192.

Thurman, R.E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M.T., Haugen, E., Sheffield, N.C., Stergachis, A.B., Wang, H., Vernot, B., et al. (2012). The accessible chromatin landscape of the human genome. Nature *489*, 75–82.

Tillo, D., and Hughes, T.R. (2009). G+C content dominates intrinsic nucleosome occupancy. BMC Bioinformatics *10*, 442.

Tillo, D., Kaplan, N., Moore, I.K., Fondufe-Mittendorf, Y., Gossett, A.J., Field, Y., Lieb, J.D., Widom, J., Segal, E., and Hughes, T.R. (2010). High nucleosome occupancy is encoded at human regulatory sequences. PLoS ONE *5*, e9129.

Valouev, A., Johnson, S.M., Boyd, S.D., Smith, C.L., Fire, A.Z., and Sidow, A. (2011). Determinants of nucleosome organization in primary human cells. Nature *474*, 516–520.

van der Heijden, T., van Vugt, J.J., Logie, C., and van Noort, J. (2012). Sequence-based prediction of single nucleosome positioning and genome-wide nucleosome occupancy. Proc. Natl. Acad. Sci. USA *109*, E2514–E2522.

Visel, A., Blow, M.J., Li, Z., Zhang, T., Akiyama, J.A., Holt, A., Plajzer-Frick, I., Shoukry, M., Wright, C., Chen, F., et al. (2009). ChIP-seq accurately predicts tissue-specific activity of enhancers. Nature *457*, 854–858.

Wei, G.H., Badis, G., Berger, M.F., Kivioja, T., Palin, K., Enge, M., Bonke, M., Jolma, A., Varjosalo, M., Gehrke, A.R., et al. (2010). Genome-wide analysis of ETS-family DNA-binding in vitro and in vivo. EMBO J. *29*, 2147–2160.

Yang, L., Zhou, T., Dror, I., Mathelier, A., Wasserman, W.W., Gordân, R., and Rohs, R. (2014). TFBSshape: a motif database for DNA shape features of transcription factor binding sites. Nucleic Acids Res. *42* (Database issue), D148–D155.

Zaret, K.S., and Carroll, J.S. (2011). Pioneer transcription factors: establishing competence for gene expression. Genes Dev. *25*, 2227–2241.

Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., Li, W., and Liu, X.S. (2008). Model-based analysis of ChIP-Seq (MACS). Genome Biol. *9*, R137.

Zhang, Z., Wippo, C.J., Wal, M., Ward, E., Korber, P., and Pugh, B.F. (2011). A packing mechanism for nucleosome organization reconstituted across a eukaryotic genome. Science *332*, 977–980.

Zhang, J.A., Mortazavi, A., Williams, B.A., Wold, B.J., and Rothenberg, E.V. (2012). Dynamic transformations of genome-wide epigenetic marking and transcriptional control establish T cell identity. Cell *149*, 467–482.

Zhou, T., Yang, L., Lu, Y., Dror, I., Dantas Machado, A.C., Ghane, T., Di Felice, R., and Rohs, R. (2013). DNAshape: a method for the high-throughput prediction of DNA structural features on a genomic scale. Nucleic Acids Res. *41* (Web Server issue), W56–W62.